Feng Kang Distinguished lecture

Author

Nicholas J. Higham, School of Mathematics, University of Manchester

Title

Exploiting Low Precision Arithmetic in the Solution of Linear Systems

Abstract

The landscape of scientific computing is changing, because of the growing availability and usage of low precision floating-point arithmetic, which provides advantages in speed, energy, communication costs and memory usage over single and double precisions. Of particular interest are the IEEE half precision (fp16) and bfloat16 arithmetics, the hardware support for which is primarily motivated by machine learning. Given the availability of these arithmetics, mixed precision algorithms that work in single or double precision but carry out part of a computation in half precision are now of great interest for scientific computing.

We consider solving a linear system $Ax = b$, with double precision A and b, by the use of a half precision LU or Cholesky factorization and mixed-precision iterative refinement. Among the points we discuss are

- how to avoid underflow and overflow, given the limited range of fp16 arithmetic,
- how to carry out error analysis for algorithms that use fp16 or bfloat16 arithmetic,
- how to simulate low precision arithmetic when hardware implementations are not available,
- the attainable speedups over state of the art solvers on current GPUs.